(The following material was taken from the appendix of Précis moderne cryptographie (Paris, 1953) by Charles Eyraud. The translation from French to English was make by Steven M. Taylor.)

C. Simple examples of the application of the calculation of probabilities in cryptography.

1. Usefulness of such calculations; their value and accuracy.

One would have to very naive to assume that any calculations, even very long and difficult ones, will permit of decryptment. The usefulness of probability calculation in cryptography is undeniable; nevertheless, it will always remain rather limited in scope. In fact, one can never really determine the value of any given secret symbol by calculation; rather, one sometimes hardly arrives at anything but the solution of certain problems of the probability of causes.

When the decryptor does not know how a cryptogram has been obtained, he can distinguish some occurrence-effects, such as, for example, the repetition of a cryptographic sign. Then, by forming various hypotheses as to the means of encipherment, and by attributing to thes occurrence-causes suitable a priori probabilities, he will be able to calculate the probability of the occurrence-effect for each hypothesis. Sometimes this will be enough to confirm his ideas; sometimes it will necessary to have recourse to Bayes' formula. But even in this event one will only obtainvaluable clues or information if the decryptor has selected the proper hypothesis with regard to the means of encipherment employed; if not, then one of the other hypotheses envisioned will necessarily seem the best.

It is therefore essential that the cryptanalyst, prior to any calculation, determine the true cause of the occurrence-effects which he observes.

Now, except in very simple cases (for example, if the frequencies of the clear text letters are retained), this first step will, in general, be difficult. In addition, even after calculation, the true cause in question will never become absolutely certain; it will become, at the most, morally certain, if the probability is close enough to 1, or absolute certainty.

"This moral certainty has shades of meaning which absolute certainty does not allow." (Quoted from M. Dumas and P. Maheu: Statistical Methods and their Uses, second installment, 1948, in the French Artillery Gazette.)

Finally, the statistical data from the observed phenomena are very rarely sufficient to permit the bringing into play the law of large numbers. Rohrbach writes: "the asymptotic laws for the calculation of probabilities (approaching the limits prescribed by the Laplace type), which are generally the most interesting part of the calculation, ordinarily play only a very secondary role in cryptography."

It follows from this that one will have to expect a considerable amount of deviation, and hence it would be almost useless to press for too much exactitude in one's calculations.

2. Unitary probabilities in cryptography.

In cryptography one is dealing with the units which are to be enciphered and those which are used to encipher the former, which are: either isolated letters or digits, or groups of letters or digits. The units to be enciphered all have their own probability of occurrence which is deduced from frequency tables (of letters, digraphs, etc.) established by statistical surveys.

The cipher units have either the same probabilities as the pure units of encipherment which each represents, or else a probability due to chance, if one puts to one side the preceding "possible cause". The "occurrence-effects" observed in the cryptogram can, moreover, be of two kinds, which vary somewhat. Either a simple occurrence will be the appearance of a given unit (clear or cipher) that one has established in advance; or else it will be the "coincidence" between two letters or groups (but not specified in advance) in texts aligned one beneath the other.

It is evident that if chance is the only factor, the probability of the appearance of a given letter is 1/26 (which we will express as 26^{-1}), and that of a numeral is 1/10 (which we will write 10^{-1}). The probabilities of this letter or figure not appearing are 25×26^{-1} and $9\times10^{-1}(= 0.9)$.

It is intuitive that one considers groups of letters or figures due to pure chance, the unitary probability of each one of them appearing is (according to the case) 26^{-2}, 26^{-3} etc., or 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} for groups of two, three, four or more letters or figures. Moreover, this is only one result of the theory of compound probabilities applied to a case in which the probability of the compound occurrences does not vary with the production of the first one among them. The probability of a group of r elements will be 26^{-r} or 10^{-r}.

It should be noticed that in a hypothetical case of pure chance, the probability of having a coincidence (not arranged in advance) is the same as having a given letter or group decided in advance. To be convinced of this, one need only think of the fact that a coincidence will occur when the letter, figure, or group agreed upon in the first text is found in the second.

3. Cases of units being enciphered. - The frequencies of units of encipherment being only averages which vary according to the authors or texts studied, the unit probabilities of appearance will never be strictly known. And when one has no need for great accuracy in caculations it is possible to accept approximate values. For example, for clear text E in French the probability is p=0.177 (whence q = 1-p = 0.823); for S, p=0.083; for A, p=0.076; etc. But it would often be sufficient to take E as being p=0.18; 0.07 for the letters SARINTULO; 0.04 for DCMP; and 0.01 for the other letters present in the text.

On the other hand, since the frequency varies with each letter considered, the probability of coincidence (not arranged in advance) will be obtained by applying the theorem of total probabilities and by adding the probabilities of coincidence appropriate to each clear (or transposed letter), themselves being equal to the square of the prbbabilities of appearance (or frequencies).

The same will be true for the clear text digraphs, from which the probability of coincidence is obtained by adding the probabilities of coincidence appropriate to each of them. But if one is dealing with digraphs formed from transposed clear letters, one must, on the contrary, use the theory of compound probabilities, taking the square of the probability of coincidence of the plain text letters.

Finally, one will have this (in formulae 25) the various unitary probabilities (in French) for letters or digraphs taken from clear, transposed, or perfectly enciphered texts with perfectly flattened cipher frequencies.

	Cryptogram	Transposition	Plain text
Coincidence probability of single letters	$\frac{1}{26} \neq \frac{4}{100}$	$\frac{7,96}{100} \neq \frac{2}{26}$	$\frac{7,96}{100} \neq \frac{2}{26} \neq \frac{8}{100}$
Coincidence probability of digraphs	$\frac{1}{26^2} \neq \frac{1,5}{1,000}$	$\frac{6,34}{1,000} \neq \frac{4,3}{26^2}$	$\frac{9,64}{1,000} \neq \frac{6,25}{26^2} \neq \frac{1}{100}$

In most cases the approximate (approached) values, whose most precise expression we have caused to follow, will suffice for the exactitude sought for in calculations. They will be adequate, for example, to indicate the fact that the probability of coincidence between plain text digraphs is clearly greater than that between digraphs formed of transposed letters.

Unit probabilities of coincidence for trigraphs and polygraphs are found to be even more different. No frequency tables exist for polygraphs which permit doing exact work, but, admitting what we have said as to the accuracy necessary for current needs, one can be content in most cases with a very simple result which we obtained by approximate reasoning starting from the frequencies of words. The coincidence of two groups of r letters in succession, whose unitary probability is 26^{-R} is the hypothetical case of pure chance, as we have indicated in paragraph 2, has a unitary probability of approximately 10^{-R} if one is dealing with successive letters in a French plain text; and, presumably, the same is true for all languages using the greco-latin alphabet. It will be noticed that this general simple law is well proven in the case of digraphs.

4. Cases of code groups. - When one studies statistically the digraphs and polygraphs of any kind of text, it is almost always preferable to consider that there are as many digraphs or polygraphs as successive letters (except the last) rather than to make any particular division which would be highly arbitrary and which would split, without a sufficiently valid reason, the most significant polygraphs.

This is no longer true in the case of code groups having four or five figures or letters. It goes without saying that even if the code is re-enciphered (except by transposition) the only valid breakdown is the one which corresponds to the individual code groups of four or five signs. As long as one is not bound by this first point, it is appropriate to consider both hypotheses.

In any event, if the code in unknown, one will not even know a priori which are the most frequent groups. But, being given what we know in general with regard to the relatively small number of frequent and very frequent code groups, one can be sure in advance that the probability of coincidence of the single code groups is much higher than what would be due to pure chance. In fact, for the most well known four-figure code (the Sittler), the two most frequent groups (punctuation signs on lines 25 and 31 of page 68) have frequencies close to $\frac{15}{1000}$, which is 150 times greater than the theoretical frequency due to pure chance. And the sum of the probabilities of coincidence of the 57 most frequent groups is already 27 times greater that the probability of coincidence between groups of four figures due to pure chance, so that the total probability of coincidence certainly exceeds by at least 30 times the probability in question.

Although the Sittler has not yet been too closely studied from this point of view, one can admit that the ratio between the theoretical unitary probabilites (due to pure chance) and the practical ones (due to code frequencies) will always be of this order of magnitude, scarcely less in the case of certain four-figure codes, and certainly more apparent for codes of five figures or groups of letters - whose theoretical frequency is much less, whereas the number of frequently appearing code groups remains very limited. In practice one can therefore take the above figures as a point of departure until such time as research in progress on the studied code permits greater accuracy.

5. The problem of repetitions. - We have defined a coincidence as being the meeting (simultaneous occurrence) of two identical elements (not specified in advance) forming a test-pair (couple). Evidently, one can have a series of successive pairs formed; for example, from elements placed in the same horizontal line in two texts which one compares by writing one beneath the other. But if one thus compares more than two texts, of if one examines statistically a single text to see how many times each element is repeated, one would expose oneself to miscalculation by using the probability of coincidence which we have just computed for test cases which are not independent of one another.

The problem which is then posed is one of the most common in cryptography, and unfortunately one of the most difficult from the mathematical point of view. Namely, that of repetitions: What is the probability of that, after N tests based upon A possible events and producing necessarily only one of them at a time, one has repetitions of events (not specified in advance) among the N results? The Chief Engineer of Naval Artillery, M. Dumas, has kindly consented to develope this problem for us in the case where the A possible events all have the same probability (1/A).

He has done it directly (see B-1), that is to say, by determining the relationship between the number of favorable cases and that of all of the possible cases.

The number of possible cases is evidently that of the number of combinations with repetitions of A elements taken N to the N: C^{N}_{A+N-1}

The number of cases in which there is no repetition is: C^{N}_{A}

The number of cases in which there is one pair and only one pair is: $A\ C^{N-2}_{A-1}$

The number of cases in which there is one three of a kind and one is: $A\ C^{N-3}_{A-1}$

The number of cases in which there is one and only one four of a kind is: $A\ C^{N-4}_{A-1}$

The number of cases in which there is two pairs and only two pairs is: $C^{2}_{A}\ C^{N-4}_{A-2}$

The number of cases in which there is three and only three pairs is: $C^{3}_{A}\ C^{N-6}_{A-3}$

The number of cases in which there are two and only two three of a kind is: $C^{2}_{A}\ C^{N-6}_{A-2}$

The number of cases in which there is one and only one full house is: $A(A-1)C^{N-5}_{A-2}$

In more general terms, the number of favorable cases for a result composed of N_1 isolated elements, N_2 pairs, N_3 three of a kind.....and eventually K elements which are all identical ($N_K = 1$) is:

$$\frac{(A+1-M-N_1)\ (A+1-M-N_1-1)\ (A+1-M-N_1-2)\ \cdots\cdots\ (A-1)\ A}{N_1!\times N_2!\times N_3!\ \cdots\cdots\times N_K!}$$

with the conventions: $M = N_2 + N_3 + \cdots\cdots + N_K$ (N_1 is excluded from this sum) and $0!=1$; $N = N_1 + 2N_2 + 3N_3 + \cdots\cdots + KN_K$

These formulae, to which we will give as a group the reference number (26), will permit us to calculate the probability of any observed repetitions. The probability P_o under which no repetition occurs, will be:

$$P_o = \frac{A!}{(A-N)!}\times\frac{(A+1)!}{(A+N-1)!} = \frac{(A-N+1)\ (A-N+2)\ \cdots\cdots\ (A-1)}{(A+1)\ (A+2)\ \cdots\cdots\ (A+N-1)}$$

The probability of having a pair will be obtained by multiplying P_o by: $\frac{N(N-1)}{A-N+1}$

The probability of having one three of a kind will be obtained by multiplying P_o by: $\frac{N(N-1)(N-2)}{(A-N+1)(A-N+2)}$

The probability of having one four of a kind will be obtained by multiplying P_o by: $\frac{N(N-1)(N-2)(N-3)}{(A-N+1)(A-N+2)(A-N+3)}$

The probability of having a full house will be obtained by multiplying P_o by: $\frac{N(N-1)(N-2)(N-3)(N-4)}{(A-N+1)(A-N+2)(A-N+3)}$

M. Dumas also observed that in general P_o is close to .5 when A is taken as approximately $N(N-1) = .7A$.

6. Newton's Formula. - It is sometimes possible to reason differntly by means of formulae (13) and (16). In fact, if in formula (13), $P_{A \sim B} = P_A + P_B - P_{A+B}$, P_{A+B} is small encugh to be neglected, the sum of the probabilities of A and B will give an approximate value in excess of the probability that one obtain A or B. And this can be extended nearer and nearer the probability that one have A or B or C. But this type of reasoning will be found to be most interesting in those special cases(which can be important) where P_{A+B} is a null.

For example, each time that M is larger than (N-M) in formula (16) it is impossible that one simultaneously have MA and MB in the course of N trials. One will therefore have $P_{A \sim B}$ by the addition of P_A and P_B. By generalizing little by little, one can therefore obtain, by simple addition, the total probability of having any given one of the A possible events, not specified in advance, at the time of each of the N trials, provided that $2M > N$.

If each of the A events (A, B. C, etc.) is equally probable, the addition of the A terms can be replaced in a multiplication by A (=1/P) of the probability P_N^M, whence:

$$(27) \quad E_N^M = A\, P_N^M = \frac{1}{P} P_N^M = \frac{N!}{M!\,(N-M)} P^{M-1} Q^{N-M}$$

By applying this formula one can, for example, calculate the probability of finding, due to pure chance, two identical letters in a trigraph of random text; then in a trigraph formed of clear transposed letters. One finds 3×0.037 in the first case and about twice that in the second case.

It goes without saying that one would leave oneself liable to serious errors by using formula (27) to study any total probability outside of the stated conditions. That is because this formula, called Newton's formula, does not, in general, furnish a probability, but rather a mathematical expectation. Rohrbach (already quoted) used it to calculate "the number of groups that one must expect to repeat M times according to the laws of chance in a list or aggregation of N groups".

We extract the following few results from curves set up by this author. For groups of five figures one will have approximately one fortuitous double repetition in the first 420 groups examined; two for 600 groups; three for 800 groups; four for 900 groups, and six for 1100 groups. The first fortuitous triple repetition should be expected in the first 4,000 groups examined.

If one considers four-figure groups, the first triple repetition is to be expected after examining only 1,000 groups. As for double repetitions, one would already occur after 135 groups, two for 200 groups, five for 300 groups, eight for 400 groups, 12 for 500 groups, and 45 for 1,000 groups.

It is also possible to calculate the mathematical expectation of there being at least one repetition in the following manner. The mathematical expectation of the first group being found among the (N-1) following groups is evidently (N-1)P; that of the second group being found among the following (N-2) groups is (N-2)P, and so on. Thus the total mathematical expectation will be:

(28) $$E = (N-1)\,P + (N-2)\,P + \cdots\cdots + P = \frac{N(N-1)}{2}\,P$$

	For groups of five figures:				For groups of four figures:			
	N	E	N	E				
One finds:	101	0.05	301	0.45	N	E	N	E
	201	0.20	401	0.80	51	0.127	101	0.505
	And E > 1		If N > 441		And E > 1		If N > 141	

Remarks: (1) Formula (27) gives in the case of 2M > N the total probability, so that any given one of A possible events, not specified in advance, is produced M times, this being true whatever the division of the A-1 other events in the N-M trials. Each particular division would, on the contrary, require a special application of the formulae (26); and it would be the sum of the partial probabilities thru arrived at which would be equal to the result furnished by formula (27).

(2) Similarly, in general, if one makes M=2, formula (27) gives the total mathematical expectation relative only to double repetitions, those repetitions of a higher order not occurring. On the contrary, formula (28) causes these higher order repetitions to occur by counting; for example, three double repetitions as a triple repetition.

These two remarks being kept in mind, the results obtained are entirely in agreement, as the reader can verify above, for the mathematical expectation of one repetition.

Finally, all of the preceding formulae assume that the A possible events all have the same probability (P = 1/A) at the time of each trial. If this condition is not satisfied, the problems of repetitions become very difficult to deal with by the calculation of probabilities; and we don't know any practical formulae which permit solving it, except in a few exceptional cases (very limited number of trials, etc.).

7. First Practical Application. What is the probability that, after having examined a large number N of cryptographic signs, one will be repeated more often than the others?

The answer is furnished immediately by Borel's rule, provided that N is large enough. Let us suppose that we have 450 cryptograms, each having an average of 40 groups of five figures (total, 90,000 figures). If we single out these figures one by one, the law of great numbers permits us to predict that each of them will be repeated about 9,000 times. The decimal unit of error is 180. $\sqrt{N} = 300$, $\sqrt{PQ} = \sqrt{0.1 \times 0.9} = \sqrt{0.09} = 0.3$, $2\sqrt{NPQ} = 2 \times 300 \times 0.3 = 180$.

In order that a figure (determined in advance) be repeated in the total of the 450 cryptograms, one has these probabilities:

Less than	or more than	x'	$Px' = 10^{-x'^2}$
8,820 times	9,180	1	10^{-1}
8,640	9,360	2	10^{-4}
8,460	9,540	3	10^{-9}
8,280	9,720	4	10^{-16}

As soon as x' is greater than 2 (or even equal to 2) the probability relative to a figure fixed in advance is so slight that one can multiply it by 10 and identify the mathematical expectation thus obtained with the probability required, so that any given figure (not specified in advance) be found, for example, less than 8,280 or more than 9,720 times in the messages.

Such a result will be due to pure chance only ONE time out of a trillion ($10^{15} = 1,000,000,000,000,000$). Practically: never. It will remain for the decipher to determine the cause of such a phenomenon. (If a code text were re-enciphered by transposition, one figure common to two or three very frequent groups would suffice to cause such an anomaly, absolutely inexplicable by chance. For example, if the 450 cryptograms were the result of such a reciphering of the Stitler, the figures PA common to the two punctuation groups which we have talked about would correspond to $x' = 3.3$ ($x' \neq 11$).)

But it can be seen that only a correct calculation permits the appreciation of its importance; and, as Emile Borel says, the result of this calculation "overwhelms the imagination".

If one were dealing with a clear text (in French and assuming the frequency of E_p to be that given by Baudouin), after having examined 90,000 letters, one will find on the average 15,930 E's ($NP = 90{,}000 \times 0.177 = 15{,}930$). The decimal unit of error being 228, one will find occurrence of it one time out of one billion to the extent of less than 15,246 or more than 16,614. But if one is dealing with cryptographic letters (with flattened frequency equal to 1/26), N=90,000 would still not be a sufficiently large number according to Mr. Lagrange's rule. In any event, in practice, even if this rule is not respected, one can almost always be certain that a variation greater than three units of error is due to a cause other than chance.

8. Second Practical Problem. What is the probability that a given letter, for example E_p, occurs M times in a clear or transposed text of N total letters?

If N is not too large one must use the general formula (16) for each value of M. And since these various eventualities mutually exclude one another, one will be able to apply the theorum of total probabilities in order to obtain the probability relative to M included between certain values.

For example, with N=10 and P=0.17 (normal frequency of E_p), one would find: $P_0 = .142$; $P_1 = .306$; $P_2 = .296$; $P_3 = .17$; $P_4 = .064$; $P_5 = .016$; $P_6 = .003$. And the total probability so that M be included between 0 and 6 is $P = .997$. Therefore, the probability that M be included between 7 and 10 is $P' = .003$.

It is appropriate to point out that these probabilities are relative to ten letters taken by chance in a clear text (for example, ten consecutive letters in a transposition). They would be slightly different for ten successive letters in the message. In fact, one will never have ten consecutive clear text E's, but one can have ten transposed E's in a row. But except for a few limited cases of this type, there is no need to make any distinction.

Moreover, the rule $K \simeq (N+1)P$ permits the calculation for the value K for M which will give the probability maxima; determining then this last P K, starting from which one can easily obtain the neighboring values. Thus one finds, with the accuracy which the use of the slide rule entails, the probabilities set forth in the following table for various values of N and for those of M included between (K-5) and (K+5).

Table of probabilities so that there be M letters e (or E) in a clear or transposed text of N letters, for the values of M giving the highest probabilities: (P_N^K = probability maxima, and and P_N^M for M included between K-5 and K+5). The table was established on the unitary probability P = 0.177.

		Probability of M =										From K-5 to K+5
N	K	K-5	K-4	K-3	K-2	K-1	K	K+1	K+2	K+3	K+4	K+5	Total
10	1					.142	.306	.296	.170	.064	.016	.003	.997
20	3			.021	.087	.179	.230	.210	.144	.071	.031	.024	.997
30	5	.002	.018	.058	.117	.170	.190	.170	.125	.007	.041	.018	.986
40	7	.015	.040	.081	.125	.152	.164	.145	.111	.074	.043	.023	.974
50	9	.028	.057	.092	.124	.147	.148	.130	.104	.071	.044	.025	.970
60	10	.021	.042	.070	.099	.128	.135	.132	.115	.096	.069	.045	.935
70	12	.030	.051	.076	.101	.117	.124	.119	.104	.084	.062	.042	.910
80	14	.034	.052	.082	.101	.114	.117	.110	.096	.078	.058	.041	.883
90	16	.046	.065	.084	.100	.109	.110	.102	.090	.073	.056	.040	.875
100	17	.035	.051	.069	.085	.097	.104	.105	.096	.083	1077	.060	.860
110	19	.041	.056	.072	.086	.095	.099	.097	.090	.079	.064	.050	.829
150	26	.045	.056	.067	.076	.083	.085	.084	.081	.074	.064	.054	.769
200	35	.047	.055	.063	.069	.072	.074	.073	.069	.064	.057	.050	.692

Example: $P_{150}^{25} = 0.083$ (N = 150, K = 26, therefore 25 = K-1).

9. Third Practical Problem. What is the probability that in a clear (or transposed) text the most frequent letter is E? Or further, which amounts to the same thing, what is the probability that in a text enciphered by monoalphabetic substitution the most frequent letter really stands for E?

Under its second form it is immediately apparent that this problem is a question of the probability of causes, and that it is therefore necessary to apply Bayes' formula. Let us suppose that in a total of N letters the most common is repeated M times. Apriori there are 26 possible causes for these M repetitions: X=A, X=B, X=C ..... X=Z (X being the cryptographic letter occurring M times). And, a priori, before the M repetiions has been noticed, each of these causes has a probability equal to the particular frequency of the letter when considered in a clear text.

Moreover, the probability that each cause gives the established frequency when that cause itself is acting, is furnished by the general formula (16) applied by the frequency Fe, or the frequency Fq, etc., and always with the same values for M and N. Finally, Bayes's formula can be written:

$$P_E = \frac{P_e \pi_e}{P_a \pi_a + P_b \pi_b + P_c \pi_c + \cdots + P_z \pi_z} = \frac{1}{1+A}$$

$$A = \frac{P_a \pi_a}{P_e \pi_e} + \frac{P_b \pi_b}{P_e \pi_e} + \cdots + \frac{P_z \pi_z}{P_e \pi_e}$$

with each of the terms of the sum A being a ratio such as:

$$\frac{P_x \pi_x}{P_e \pi_e} = \left(\frac{F_x}{F_e}\right)^{M+1} \left(\frac{1-F_x}{1-F_e}\right)^{N-M}$$

$\pi e = Fe$, $\pi a = Fa$, etc.

In fact, in this ratio the N!, M! and (N-M)! of formula (16) disappear, and the frequency of each letter is affected only by the exponent (M+1) instead of M in order to take into account the a priori probability.

If one wants the probability Pe to be greater than X%, it is sufficient that $100 > X(1+A)$; or further, that $XA < (100-X)$. For example, to have the probability Pe greater than 50% it is sufficient to have $A < 1$. And one can calculate the value for M which is necessary so that this inequality is satisfied for various values of N.

It is evidently useless to conduct one's calculations with great accuracy, since M must be an integer and it will therefore be necessary to round off each time to the higher whole number. Consequently, one can neglect the uncommon letters which will have little effect on the result, and accept a frequency equal to the total of the letters SARINTULO. By taking 7%, which is a little high for the letters ULO, one will in practice compensate for the error introduced by neglecting the rare letters. Under these conditions we have:

$$A = 9\left(\frac{7}{18}\right)^{M+1} \quad \frac{93}{82}^{N-M}$$

It is this formula which has enabled us to establish, in a rapid manner, the following table, which gives for each value of N the M minimum number of repetitions of the most frequent letter corresponding to the values of Pe.

The following table gives the probability Pe that the most frequent letter in a monoalphabetic substitution, occurring M times in a total of N letters, represents plain text E.

N	M									K
10	3	3	4	4	5	6	6	7	7	1
20	4	5	5	5	6	7	7	8	8	3
30	5	6	6	7	7	8	8	9	10	5
40	6	7	7	8	8	9	10	10	11	7
50	7	8	8	9	10	10	11	11	12	9
60	8	9	10	10	11	11	12	12	13	10
70	10	10	11	11	12	13	13	14	14	12
80	11	12	12	13	13	14	14	15	15	14
90	12	13	13	14	14	15	15	16	16	15
100	13	14	14	15	15	16	16	17	18	17
110	14	15	16	16	17	17	18	18	19	19
150	19	20	20	21	21	22	22	23	24	26
200	25	26	26	27	27	28	28	29	30	35
Pe=	50%	66%	75%	80%	90%	95%	97%	98%	99%	

For example, if the most frequent letter of a text totalling 50 letters, enciphered monoalphabetically, is repeated ten times, there are 95 chances out of 100 that it is clear text E; but if this most frequent letter is repeated only eight times there are but 75 chances out of 100 that it will be the clear E.

It will be noticed that in clear texts of 50 letters there are, on the average, nine E's ($K = 50 \times .177 = 9$). Moreover, there will be fewer than nine E's almost as often than there will be more than nine E's. The table on the preceding page easily permits us to verify this point, because it gives the exact indication of K-5 to K+5; and one can assume an equal division for the last values greater and lesser than the preceding ones.

In the case of texts having 30 letters (K=5) there will be a chance of better than one out of two ($Pe > 50\%$) of recognising E, by its greater frequency only if M is greater than K ($M > 5$), which is only one time out of two. Now it is this problem which would confront a decryptor who had previously succeeded in making columns of letters each enciphered by the same key symbol.

It is therefore evident that if these columns are less than 30 letters in height, the decryptor of a multiple alphabet procedure will have very little chance of success from this single base of departure. But in fact it will never be alone; many other small characteristics will yield help in guessing a probable word, which would completely alter the problem of probabilities.

For texts of 110 letters (K=19) the clear E, on the contrary, will almost certainly be indicted by its frequency (Pe=99%) about one time out of two. Finally, for text of 200 letters the total probability of E_p included between 30 and 40 being .692 (see the preceding table), one can say that 85 times out of 100 (.69 + .16 = .85) plain text E will be indicated by its higher frequency with almost complete certitude (Pe=99%). And it would always be indicated in actual practice with a probability greater than 50% because one will practically never have fewer than 25 E's in 200 clear text letters.

10. Probability of frequencies in transposition.

If one has established from statistical samples the normal frequencies in a transposition, there is evidently no difficulty in calculating by formula (15) of the generalized alternative, the probability of the total event, since the theoretical frequencies of the clear text letters are known. Moreover, the probability in question will in general be very slight in absolute value, and even more slight as N is greater. On the contrary, it will be very large compared to the probability of the general distribution occurring in the event of pure chance; and in general proportionally larger as N increases. This is due to the fact that the law of great numbers and everything which we have said in B-10 about relative and absolute deviations remains true for a generalized alternative.

And in practice one very quickly attains this "moral certitude" of which Messieurs M. Dumas and P. Maheu speak, since for 100 clear text letters, by taking into account only the letters ESARINTULO, the probability furnished by the hypothesis of pure chance will be on the AVERAGE 10^{25} times less than that given by the correct hypothesis.. In other words, one would have to make 10^{25} attempts in order to have the mathematical expectation that one of these attempts give by chance the same statistical results as a particular simple transposition (chosen in advance as a standard).

In order to give the reader an idea of a number equal to or greater than 10^{25}, we will seek inspiration from an example given by Emile Borel. If each inhabitant of the globe made 100 trials of this kind per second for a thousand centuries, the total number of trials would still be less than 10^{25}. Thus one sees that except in the case of an unknown third cause, one has a practically absolute certainty that one is really dealing with a transposition, even if this means of encipherment had seemed unlikely a priori.

Note: What we have just said clearly indicates that the problem dealt with was, in reality, a matter of probability of causes (or as some say, of inverse probabilities), and that one could make use of Bayes' formula. But in the actual occurrence it would have been very difficult to choose a priori probabilities for each of the two causes envisaged; and if one supposes them to be equal to one another, Bayes' formula lacks interest, since it amounts to comparing directly, as we have done, the two a posteriori probabilities.

11. Frequential probabilities in monoalphabetic substitution.

If the problems of the calculations of probabilities which are raised with regard to transpositions are infrequent (almost uniquely those we have dealt with in sections 8, 9, and 10), and are easy to solve, this is no longer at all true in the case of monoalphabetic substitution.

The ordinary simple substitution of letters already sometimes presents insurmountable difficulties from this point of view, of which the problem of repetitions (C-5) has given us a sample. On the other hand, the calculation of probabilities will in general be of more help in solving substitution than in solving transposition. For example, under certain conditions it will enable us to choose between different possible substitutional keys, or at least to classify them by order of increasing probability, which never happens with transpositional keys.

In fact, if one knows in advance that one is dealing with a normally arranged substitution alphabet, or even one parallel to a known base alphabet, or even being a part of any given known family, one will be able to perform the calculation 26 times by applying formula (15) of the generalized alternative to each one of the 26 hypotheses. Since the latter are all equally probable a priori, it will not even be necessary to use Bayes' formula to choose between them unless one is anxious to have a numerical expression of their relative a posteriori probabilities.

This numerical expression is, moreover, very exact. We have performed the calculation in the case of alphabets shifted only with regard to the order of theoretical frequencies (alphabets ESARINTULODC.....) for 30 clear letters: 5 E's, 3 A's, 3 S's, 3 R's, 2 I's, 2N's, 2 T's, 2 U's, 2 L's, 2 O's, 1 D, 1 C, 1 M, 1 P.

We found that the probability of a zero shift was greater than 999/1000, and that of a shift of one (the most probable) did not exceed 1/2000. This result satisfactorily completes the indication previously given of the faint probability of being able to recognize the clear E on the basis of its higher frequency in a column of 30 letters. It shows in fact how much the question differs in the case of known parallel alphabets (even not-normally parallel) when one can bring into play the frequencies of all of the letters.

It goes without saying that the same reasoning and the same calculations can be made eventually for more than 26 simple keys known to the decryptor. But it is evidently impossible to begin them again in the factorial 26 hypotheses which must be envisioned for a completely unknown substitution alphabet. One can, however, believe that approximate calculations will have some meaning if one conducts them in the following manner.

1. One will take approximate theoretical frequencies as has already been indicated (C-2) in such a way as to have classes of letters: E-SANTIRULO-DCMP- etc., having the same probability and which are not distinguished from one another under this point of view.

2. One will operate as we have stated in regard to transposition, attributing the frequency of E to the most frequent cryptographic letter, the frequencies of SANTIRULO to the other common letters, and on on.

3. One will begin again the calculations of other and less favorable hypotheses (for in reality the ten most frequent letters will not often be all translations of clears - ESARINTULO) so as to determine to what degree one should depend upon the results.

Another procedure would be to calculate by formula (26) the probability that pure chance has given the totality of the observed frequencies. But this probability will have no meaning by itself, and it will be necessary to compare it to those which the same formulae would give for collections containing as many letters as a whole and chosen arbitrarily either in clear texts or in other cryptographic texts.

One would be able to reason further on the probabilities of "coincidences", as we explain hereafter for polyalphabetic substitution, and taking account of the fact that these coincidences must be produced for any shift, and not for a particular shift.

As for the special problem which consists of recognizing if two collections of cryptographic letters originate from one single simple key or from two different keys, it also will have to be dealt with differently, depending upon whether one must identify totally unknown keys, or only determining keys taken from a family (or a sub-family) of known keys.

One will always be able to be content with an approximate solution by attributing arbitrarily the frequencies of ESARINTULO to the most frequent letters in the ensemble of the two collections; then by appreciating the feeling (or by comparison with the results obtained from clear texts of similar importance) if the aforementioned frequencies give satisfactory results for calculating the probabilities of both collections.

But in reality one can only deal with the problem with sufficient accuracy if one knows a priori the possible keys - provided they be limited in number. Further, it is necessary to make a double application of Bayes' formula for each double hypothesis, while choosing, first of all, the hypothesis most favorable to one collection and then to the other.

Finally, everything which we have just said would eventually be true for pairs, whether the single substitution be monographic or digraphic. But in the latter case it would evidently be necessary that the subdivision into digraphs be correct.

12. The Case of Polyalphabetic Substitution.

It will be sufficient if we remember that the characteristics of multiple alphabet substitution are, in reality, a series of characteristics of monoalphabetic substitutions - in order *not* having to repeat with regard to polyalphabetic substitution almost everything that we have just said. However, this is only true when one has succeeded in reconstituting the collections of cryptographic letters in each single alphabet. And the search for cipher texts which are comparable with one another can lead to special problems of probability calculation.

We know that multi-alphabetic substitution preserves, in the case of two text enciphered with the same principal key, all of the coincidences which exist between the two clear texts. The probability of these coincidences is therefore greater than that (due to pure chance) which would come into play in the case of two cryptograms enciphered with different principal keys. (In all accuracy, one should say: different arbitrary principal keys.)

In principle, nothing is simpler than applying Bayes' formula in order to obtain a <u>posteriori</u> (when one has observed the coincidences) the probability of an accurate comparison. But in practice, lacking the knowledge of how to calculate the probability of a compound event by coincidence of polygraphs, digraphs, and isolated letters, one will have to choose between reasoning on either the polygraphs, digraphs, or single letters (with unitary probabilities of 26^{-R}, 26^{-2} and 26^{-1} in the case of pure chance; and 10^{-R}, 6.25×26^{-2}, 2×26^{-1} if the comparison is good).

Theoretically, however, it would be necessary to separately study all of the possible divisions into digraphs and polygraphs.

In practice, although this method seems difficult to justify in theory, one will obtain enough good results by considering the N-1 digraphs formed with successive letters. In this work a trigraph counts for two digraphs; and this is approximately correct. This does not succeed as well for longer polygraphs, but the approximation is nevertheless acceptable.

But the principal difficulty will be, as in almost all cases when one wants to apply Bayes' Formula, in judging the a priori probabilities. For it is evident that they will vary a great deal, depending upon the total number of messages compared and on the different keys employed.

If one considers pure chance, a primary question can be asked relative to the mathematical expectation of repetitions. The formulae (27) or (28) should, it seems, permit one to easily recognize whether it is or is not normal to find repetitions of polygraphs of order R after having compared N cryptograms in all possible ways and with all possible variations. One will have performed $N(N-1)\times D/2$ comparisons of the cryptograms (D being the average number of variations or differences in value). But these will not be independent comparisons; therefore the calculation will be only approximate.

Finally, the eventual periodicity of one main key should, in principle, appear as a series of favorable results for different variations (value shifts) separated by whole multiples of the key length. In practice, aside from exceptional cases, such a series of calculations of probabilities would doubtless be an exhausting if not useless task for purposes of this observation.

13. Special Problem. Study of the Value of a Criterion.

We know (Summary II-104) that when the cryptanalyst is not acquainted with the incoherent alphabets of the Bazeries Cylinder (or an analogous system) his only recourse is to determine the length N of the cryptographic lines, then to classify these lines in 24 (or 25) families according to their similarities.

What criterion is acceptable in orddr to associate two rows; and what will be the probability of a valid association? In other words, of what value will be the criterion adopted? Let us apply Bayes' formula.

In the case of a system having 26 generatrices - one clear and 25 cipher - there is a priori one chance out of 25 that one has associated two cipher lines with the same generatrix; and 24 chances out of 25 for the contrary case. Whence:

$$P' = \frac{P'}{P' + 24P^2}$$

(P' being the probability calculated according to the index of coincidence of clear text single letters or digraphs, and P^2 the result of the same calculation on the hypotesis of chance). It is evidently necessary that P' be greater than 50% or else the criterion has no value. One must therefore have $P' > 24P^2$.

Formula (16) applied twice with 26^{-1} and 2×26^{-1} as the unitary probability of the coincidences of letters immediately gives the following, in which the N!, M!, and (N-M) disappear.

$$(2\times26^{-1})^M(1-2\times26^{-1})^{N-M} > 24(26^{-1})^M(1-26^{-1})^{N-M}$$

$$2^M(1-2\times26^{-1})^{N-M} > 24(1-26^{-1})^{N-M}$$

or further approximately: $2^M > 24[1+(N-M)26^{-1}]$

It will be noticed that the M number of coincidences of letters must be equal to at least five times whatever the number N of disks. The variations of M, in fact, have much more influence than those of N on this unequality.

A completely analogous calculation leads to the conclusion that there must be at least two coincidences of digraphs, or else one coincidence of a trigraph (equalling two digraphs) in order to associate two lines with a half-chance of success. In fact, one will have to be more demanding unless one wants to have half of the associated lines incorrect in a collection thus formed.

One should notice the relatively slight influence of the lengths of the lines (N = number of disks) on the value of the criterion, in order to deduce from it that if this length decreases, security will not be too greatly diminished. In fact, if the total number of lines increases, the total number of good associations will decrease - because there will be fewer chances of having digraphs or polygraphs in common between short lines; so that finally security will vary negligibly with the number of disks.

On the other hand, and this is what had particularly impressed our late regreted friend Saillard with regard the the special case treated by him in an article which he had devoted to this question and from which we have drawn part of our inspiration, one sees to what degree a criterion inprudently accepted in an arbitrary manner could lead to serious errors.

In reality, the decryptor should bring together lines having at least three digraphs in common (or one digraph and a trigraph .....etc.). He will rarely have enough in total so that the method analysis can bear fruit, and the system will remain undecryptable except in exceptional cases - mistakes in encipherment such as the habitual use of a small number of generatrices, etc.

14. The Case of Codes. Conclusions.

What we have said about the frequencies of code groups (in Summary III-17) and about their probability of coincidence (C-4) will permit the solving of problems relative to code substitutions by methods analogous to those which we have recommeded up to now. In reality, we have limited outselves to very simple practical applications, in cases where the calculation of probabilities will sometimes have only a slight practical interest - for, after all, one will not perform a calculation of probabilities in order to solve an ordinary simple substitution.

These ideas could really be developed in much more detail. And they doubtless will be when a practical application is wanted, notably in the case of code groups having in general slight but unknown probabilities, of the general ides discussed in the excellent study already cited of Messrs. M. Dumas and P. Mahou on the subject of: "The verification or the test of an hypothesis"; "The suppositions on a new series of results"; "The comparison of two series of results"; etc. Moreover, these authors most often recommend graphic methods, especially those of "Henry's graphic", for the solution of such problems.

Sometimes one will need to make use of the limit-laws of the Poisson type (law of small probabilities) which, according to Rohrbach, have more practical application in cryptography than those of the Laplace type. In certain cases one will fruitfully have recourse to another chapter in the Theory of Probabilities which we have not been able to deal with here - the one dealing with contingent events which cause one another according to variable probabilities depending upon their own appearance.

In fact, this is really the case of clear letters or of code groups whose probability varies with the meaning of the preceding letters or groups. These questions have been studied in detail by French mathematicians, notably Emile Borel, but especially Frechet, in particular in Methods of arbitrary functions: theory of "chained" events in the case of an infinite number of possible states (Gautheir-Villars, 1938).

But in this necessarily very condensed expose we have especially wished to give sensible foundations from which each cryptologist will be able to develope his own research. If not contained by strict rules, calculation can easily end in erroneous probabilities which sometimes push fantasy until it becomes greater than unity. We hope that our reader, duly forwarned, will at least avoid such a heresy by checking carefully, so that his formulae can never give such a result, whatever the values attributed to the various elements may be.

Here the reader has an opportunity to measure mathematically, so to speak, the difficulties which one runs into (generally without being able to measure them) in the practice of decrypting.

Unitary probabilities of coincidence between: 2 letters; 2 digraphs.

Chosen at random in texts written in	Russian:	.047	.005
	German:	.082	.011
	English:	.067	.008
	Spanish:	.077	.01
	Italian:	.075	.008
	French:	.08	.01